Student Academic Dropout Prediction

1.Objectives:

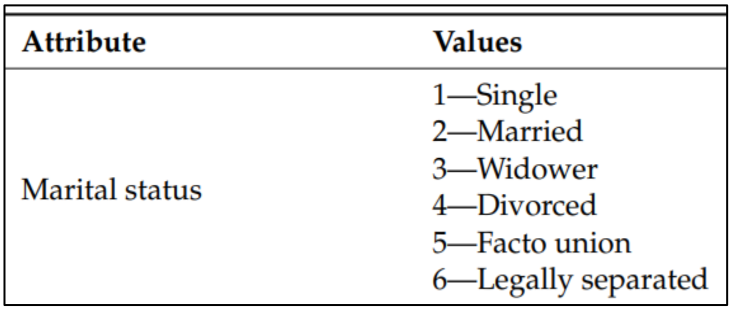
- To study historical data and understand factors related to success and dropout of students
- To study how these factors contribute to their academic performance
- To identify interdependencies and interactions between variables affecting students' overall performance and areas of improvements
- To give timely suggestions can be given to educational institutions to implement precautionary measures required to reduce attrition rate of students

2.Data:

- A second-hand dataset from Kaggle - https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention
- Data consists of 4424 records of 35 variables

3.Data Description:

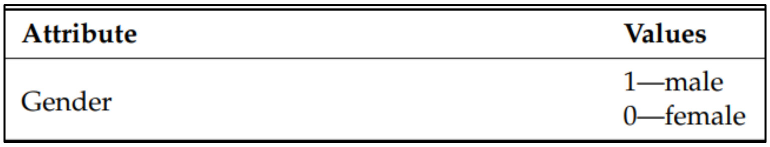
The dataset contains 35 variables which are coded as numeric. For example, Marital status is coded as numbers from 1 to 6, each with different categories as below:



Attribute	Values
Marital status	1—Single 2—Married 3—Widower 4—Divorced 5—Facto union 6—Legally separated

Figure 1: Values for attribute Marital Status

Similarly, gender variable is coded as 1 and 2, with categories as below:

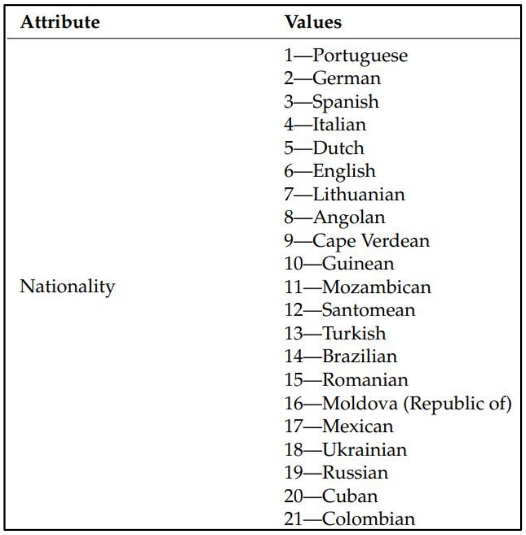


Attribute	Values
Gender	1—male 0—female

Figure 2: Values for attribute Gender

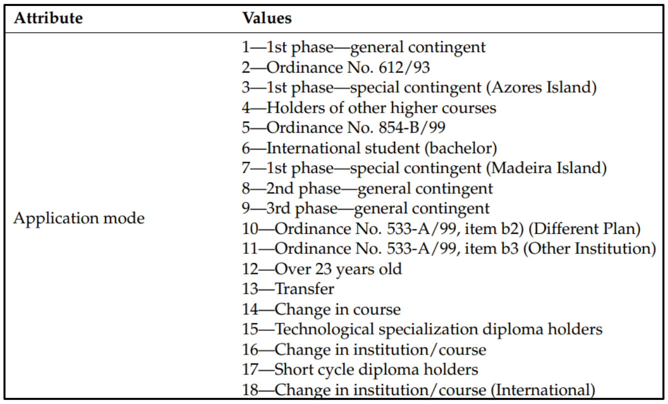
The main outcome variable is named “Target” in the dataset with categories “Dropout”, “Enrolled” and “Graduate”.

The Description of all the other coded categories is given below:



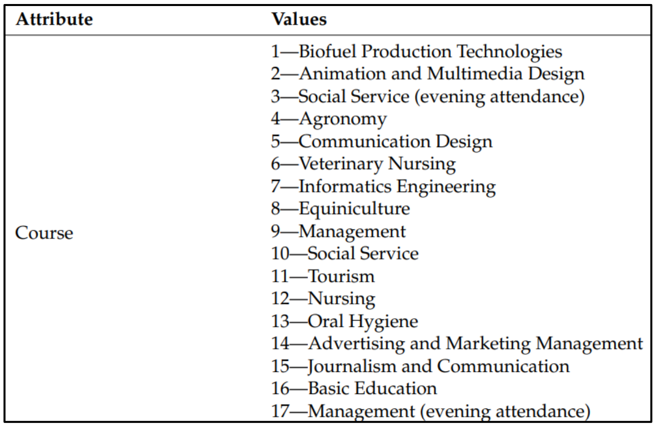
Attribute	Values
Nationality	1—Portuguese 2—German 3—Spanish 4—Italian 5—Dutch 6—English 7—Lithuanian 8—Angolan 9—Cape Verdean 10—Guinean 11—Mozambican 12—Santomean 13—Turkish 14—Brazilian 15—Romanian 16—Moldova (Republic of) 17—Mexican 18—Ukrainian 19—Russian 20—Cuban 21—Colombian

Figure 3: Values for attribute Nationality



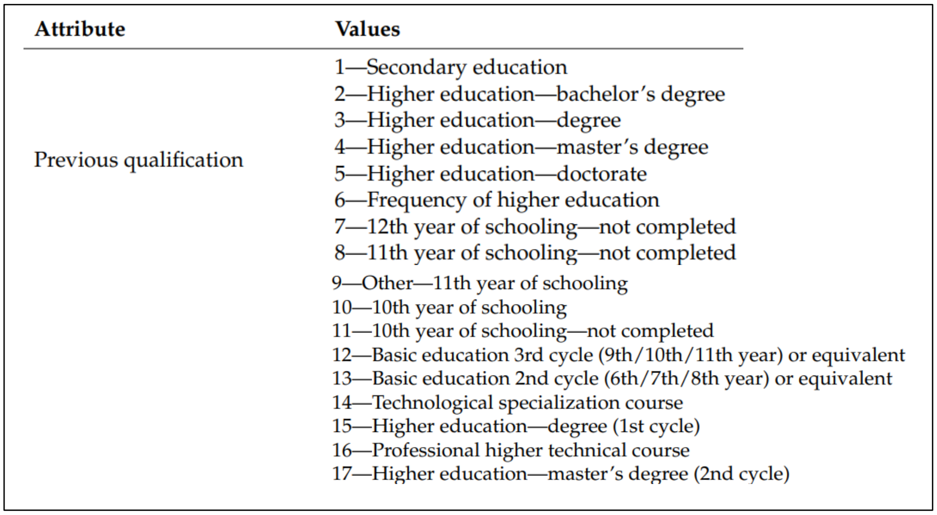
Attribute	Values
Application mode	1—1st phase—general contingent 2—Ordinance No. 612/93 3—1st phase—special contingent (Azores Island) 4—Holders of other higher courses 5—Ordinance No. 854-B/99 6—International student (bachelor) 7—1st phase—special contingent (Madeira Island) 8—2nd phase—general contingent 9—3rd phase—general contingent 10—Ordinance No. 533-A/99, item b2) (Different Plan) 11—Ordinance No. 533-A/99, item b3 (Other Institution) 12—Over 23 years old 13—Transfer 14—Change in course 15—Technological specialization diploma holders 16—Change in institution/course 17—Short cycle diploma holders 18—Change in institution/course (International)

Figure 4: Values for attribute Application mode



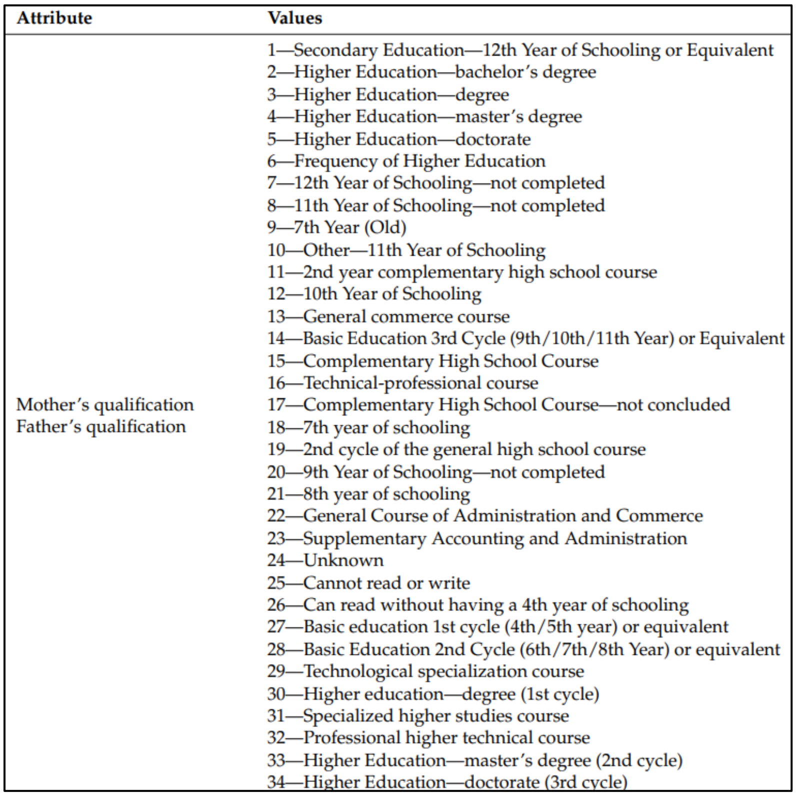
Attribute	Values
Course	1—Biofuel Production Technologies 2—Animation and Multimedia Design 3—Social Service (evening attendance) 4—Agronomy 5—Communication Design 6—Veterinary Nursing 7—Informatics Engineering 8—Equiniculture 9—Management 10—Social Service 11—Tourism 12—Nursing 13—Oral Hygiene 14—Advertising and Marketing Management 15—Journalism and Communication 16—Basic Education 17—Management (evening attendance)

Figure 5: Values for attribute Course



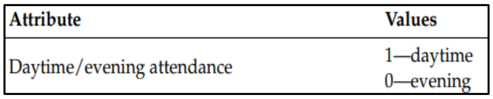
Attribute	Values
Previous qualification	1—Secondary education 2—Higher education—bachelor's degree 3—Higher education—degree 4—Higher education—master's degree 5—Higher education—doctorate 6—Frequency of higher education 7—12th year of schooling—not completed 8—11th year of schooling—not completed 9—Other—11th year of schooling 10—10th year of schooling 11—10th year of schooling—not completed 12—Basic education 3rd cycle (9th/10th/11th year) or equivalent 13—Basic education 2nd cycle (6th/7th/8th year) or equivalent 14—Technological specialization course 15—Higher education—degree (1st cycle) 16—Professional higher technical course 17—Higher education—master's degree (2nd cycle)

Figure 6: Values for attribute Previous Qualification



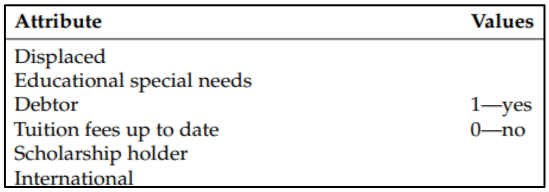
Attribute	Values
Mother's qualification Father's qualification	1—Secondary Education—12th Year of Schooling or Equivalent 2—Higher Education—bachelor's degree 3—Higher Education—degree 4—Higher Education—master's degree 5—Higher Education—doctorate 6—Frequency of Higher Education 7—12th Year of Schooling—not completed 8—11th Year of Schooling—not completed 9—7th Year (Old) 10—Other—11th Year of Schooling 11—2nd year complementary high school course 12—10th Year of Schooling 13—General commerce course 14—Basic Education 3rd Cycle (9th/10th/11th Year) or Equivalent 15—Complementary High School Course 16—Technical-professional course 17—Complementary High School Course—not concluded 18—7th year of schooling 19—2nd cycle of the general high school course 20—9th Year of Schooling—not completed 21—8th year of schooling 22—General Course of Administration and Commerce 23—Supplementary Accounting and Administration 24—Unknown 25—Cannot read or write 26—Can read without having a 4th year of schooling 27—Basic education 1st cycle (4th/5th year) or equivalent 28—Basic Education 2nd Cycle (6th/7th/8th Year) or equivalent 29—Technological specialization course 30—Higher education—degree (1st cycle) 31—Specialized higher studies course 32—Professional higher technical course 33—Higher Education—master's degree (2nd cycle) 34—Higher Education—doctorate (3rd cycle)

Figure 7: Values for attribute qualification of parents



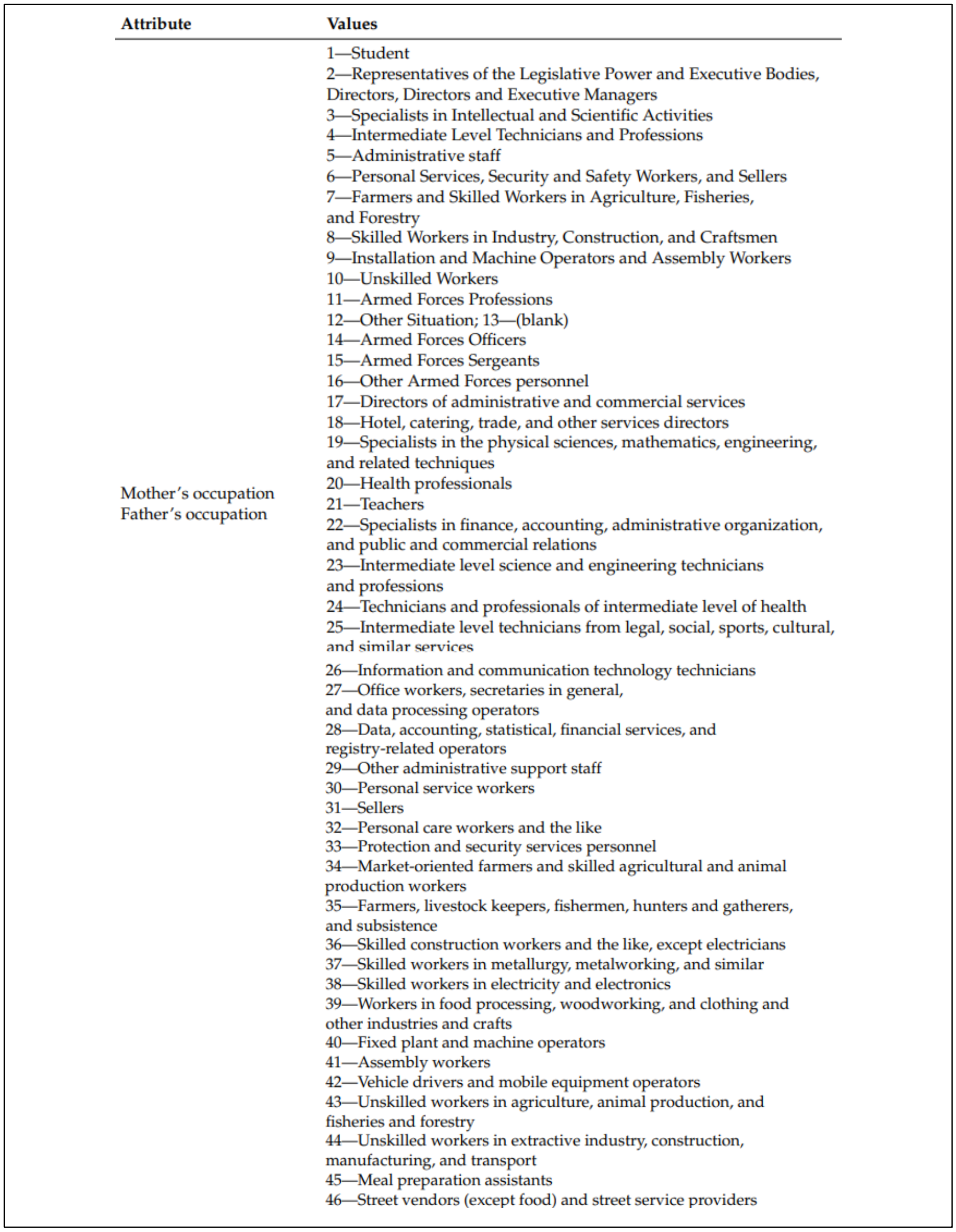
Attribute	Values
Daytime/evening attendance	1—daytime 0—evening

Figure 8: Values for attribute attendance



Attribute	Values
Displaced Educational special needs Debtor Tuition fees up to date Scholarship holder International	1—yes 0—no

Figure 9: Values for Miscellaneous attributes



Attribute	Values
Mother's occupation Father's occupation	1—Student 2—Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 3—Specialists in Intellectual and Scientific Activities 4—Intermediate Level Technicians and Professions 5—Administrative staff 6—Personal Services, Security and Safety Workers, and Sellers 7—Farmers and Skilled Workers in Agriculture, Fisheries, and Forestry 8—Skilled Workers in Industry, Construction, and Craftsmen 9—Installation and Machine Operators and Assembly Workers 10—Unskilled Workers 11—Armed Forces Professions 12—Other Situation; 13—(blank) 14—Armed Forces Officers 15—Armed Forces Sergeants 16—Other Armed Forces personnel 17—Directors of administrative and commercial services 18—Hotel, catering, trade, and other services directors 19—Specialists in the physical sciences, mathematics, engineering, and related techniques 20—Health professionals 21—Teachers 22—Specialists in finance, accounting, administrative organization, and public and commercial relations 23—Intermediate level science and engineering technicians and professions 24—Technicians and professionals of intermediate level of health 25—Intermediate level technicians from legal, social, sports, cultural, and similar services 26—Information and communication technology technicians 27—Office workers, secretaries in general, and data processing operators 28—Data, accounting, statistical, financial services, and registry-related operators 29—Other administrative support staff 30—Personal service workers 31—Sellers 32—Personal care workers and the like 33—Protection and security services personnel 34—Market-oriented farmers and skilled agricultural and animal production workers 35—Farmers, livestock keepers, fishermen, hunters and gatherers, and subsistence 36—Skilled construction workers and the like, except electricians 37—Skilled workers in metallurgy, metalworking, and similar 38—Skilled workers in electricity and electronics 39—Workers in food processing, woodworking, and clothing and other industries and crafts 40—Fixed plant and machine operators 41—Assembly workers 42—Vehicle drivers and mobile equipment operators 43—Unskilled workers in agriculture, animal production, and fisheries and forestry 44—Unskilled workers in extractive industry, construction, manufacturing, and transport 45—Meal preparation assistants 46—Street vendors (except food) and street service providers

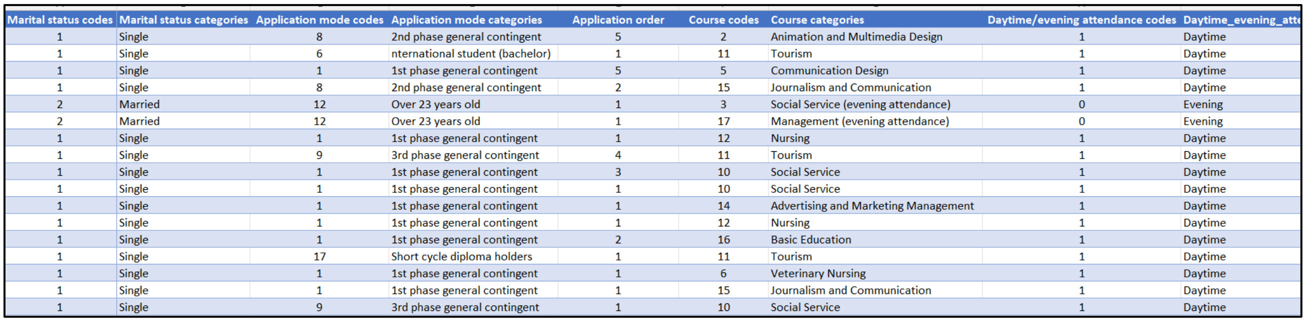
Figure 10: Values for attribute occupation of parents

4.Data Analysis:

- The data is skewed as most of the students enroll in the age group 20-30, so we have more records with age 20-30

5.Data Pre-processing and evaluating plots:

- Converting the numerical categories into nominal categories



Marital status codes	Marital status categories	Application mode codes	Application mode categories	Application order	Course codes	Course categories	Daytime/evening attendance codes	Daytime_evening_atte
1	Single	8	2nd phase general contingent	5	2	Animation and Multimedia Design	1	Daytime
1	Single	6	nternational student (bachelor)	1	11	Tourism	1	Daytime
1	Single	1	1st phase general contingent	5	5	Communication Design	1	Daytime
1	Single	8	2nd phase general contingent	2	15	Journalism and Communication	1	Daytime
2	Married	12	Over 23 years old	1	3	Social Service (evening attendance)	0	Evening
2	Married	12	Over 23 years old	1	17	Management (evening attendance)	0	Evening
1	Single	1	1st phase general contingent	1	12	Nursing	1	Daytime
1	Single	9	3rd phase general contingent	4	11	Tourism	1	Daytime
1	Single	1	1st phase general contingent	3	10	Social Service	1	Daytime
1	Single	1	1st phase general contingent	1	10	Social Service	1	Daytime
1	Single	1	1st phase general contingent	1	14	Advertising and Marketing Management	1	Daytime
1	Single	1	1st phase general contingent	1	12	Nursing	1	Daytime
1	Single	1	1st phase general contingent	2	16	Basic Education	1	Daytime
1	Single	17	Short cycle diploma holders	1	11	Tourism	1	Daytime
1	Single	1	1st phase general contingent	1	6	Veterinary Nursing	1	Daytime
1	Single	1	1st phase general contingent	1	15	Journalism and Communication	1	Daytime
1	Single	9	3rd phase general contingent	1	10	Social Service	1	Daytime

Figure 11: Conversion of numerical categories to nominal categories

- Made use of plots like Histogram and Boxplot to understand distribution of variables like age, GDP, inflation rate, gender, course categories.

6. Model Evaluation using ROC:

- The area under the Curve (AUC) is 0.94 showing that the logistic regression model built is performing equally well on training data as well as test data set.

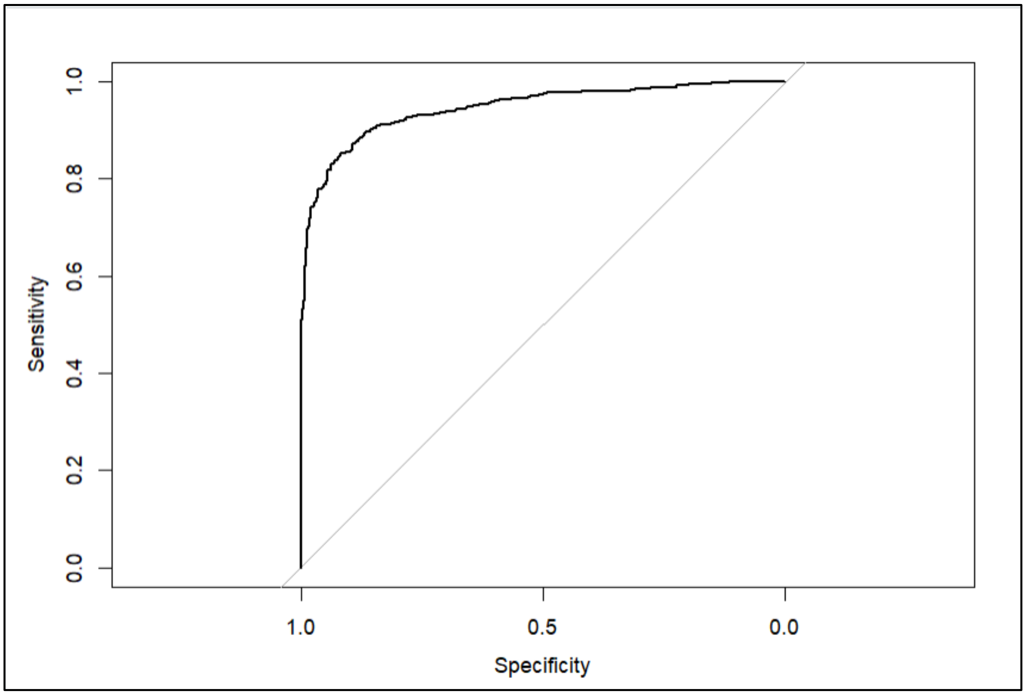


Figure 12: ROC Curve

7. Results:

From the summary of Logistic Regression Model, we can say that “Course Codes”, “Debtor Codes”, “Tuition fees up to date codes”, “Gender codes”, “Scholarship holder codes” , “Curricular units 1st sem (credited)” , “Curricular units 1st sem (enrolled)”, “Curricular units 1st sem (approved)” have a major impact on predicting the drop out probability.

8.Conclusion:

The odds of students dropping out are highly dependent on the course they are enrolled in, if they have any debts, age, gender and curricular units and grades in 1st semester.

Based on analysis, we can provide suggestions to academic institutions to reduce the dropout rate among students. Since dropout rate depends on curricular units, institutions can provide higher flexibility in course options. Students with debt may receive additional student assistance. Many students in the age group above 30 have obligations outside school such as looking after families and employment. To avoid drop out of these students, institutions can avail themselves of the facility of online classes.

9.References:

https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention
https://cran.r-project.org/
https://utdallas.primo.exlibrisgroup.com/permalink/01UT_DALLAS/2hgl0t/alma9927850104601421